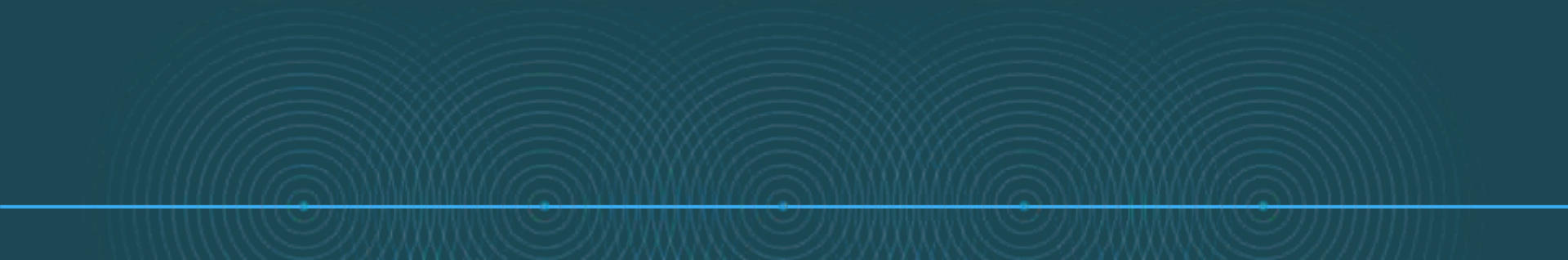


Raising the standard of published systematic reviews

A case study from chemical risk research

Paul Whaley

23 May 2018



About me

- Researcher at Lancaster University and the Evidence-Based Toxicology Collaboration at Johns Hopkins BSPH
- Background in environmental health advocacy and science communication
- Introduced to systematic reviews as gold-standard approach to evidence synthesis in early 2010
- Associate Editor for Systematic Reviews at *Environment International* (IF 7.088) – first specialist EH SR editor
- The “frameworks guy”: systematic approaches to evidence surveillance and synthesis; critical appraisal tools; codes of practice; research quality management

Today's presentation

- Reproducibility issues in chemical risk assessment as a driver of interest in systematic review methods
- Uptake of SR methods
- Challenges we are seeing (poor quality SRs)
- How we are addressing these challenges at *Environment International*
- Implications for you as potential submitting authors and conductors of systematic reviews

A “reproducibility crisis” in primary research

VIEWPOINT

The Proposal to Lower P Value Thresholds to .005

John P. A. Ioannidis, MD, PhD
 Stanford University School of Medicine
 Meta-Research Innovation Center at Stanford, Department of Medicine, Health Research and Policy, Biostatistics Center, and Department of Statistics, Stanford University, Stanford, California

P-values and accompanying methods of statistical significance testing are creating challenges in biomedical science and other disciplines. The vast majority (96%) of articles that report P values in the abstract, full text, or both include some value of .05 or less.¹ However, many of the claims that these reports highlight are likely false.² Recognizing the major importance of the statistical significance convention, the American Statistical Association (ASA) published³ a statement on P values in 2016. The stated goal is widely believed to be problematic, but how exactly to fix the problem is far more contentious. The contributions to the ASA statement also include 20 independent, accompanying commentaries outlining different aspects and prioritizing different solutions. Another large coalition of 22 methodologists recently proposed⁴ a specific, simple move: lowering the routine P-value threshold for claiming statistical significance from .05 to .005 for new discoveries. The proposal met with strong endorsement in some circles and concerns in others.

P-values are misinterpreted, overvalued, and misused. The language of the ASA statement avoids the dissection of these 3 problems. Multiple misinterpretations of P values exist, but the most common one is that they represent the probability that the studied hypothesis is true.⁵ A P value of .02 (2%) is wrongly considered to mean that the null hypothesis (eg, the drug is as effective as placebo) is 2% likely to be true and the alternative (eg, the drug is more effective than placebo) is 98% likely to be correct. Over this aspect when it is forgotten that “proper inference requires full reporting and transparency,”⁶ better-looking (smaller) P values alone do not guarantee full reporting and transparency. In fact, smaller P values may tend to selective reporting and misrepresentation. The most common misuse of the P value is to make “scientific conclusions and business or policy decisions” based on “whether a P value passes a specific threshold” even though “a P value, or statistical significance, does not measure the size of an effect or the importance of a result,” and “by itself, a P value does not provide a good measure of evidence.”⁷

fully considered how low a P value should be for a research finding to have sufficiently high chance of being true. For example, adoption of genome-wide significance thresholds ($P < 5 \times 10^{-8}$) in population genomics has made discovered associations highly replicable and these associations also appear consistently when tested in new populations. The human genome is very complex, but the extent of multiplicity of significance testing involved is known, the analysis are systematic and transparent, and a requirement for $P < 5 \times 10^{-8}$ can be cogently argued at.

However, for most other types of biomedical research, the multiplicity involved is unclear and the analysis are nonsystematic and nontransparent. For most observational exploratory research that lacks prespecified protocols and analysis plans, it is unclear how many analyses were performed and what various analytic paths were explored. Hidden multiplicity, nonsystematic exploration, and selective reporting may affect even experimental research that randomized study. Even though it is now more common to have a prespecified protocol and statistical analysis plan and registration of the trial posted on a public database, there are still substantial degrees of freedom regarding how to analyze data and outcomes and what exactly to present. In addition, many studies in contemporary clinical investigation focus on smaller benefits or risks, therefore, threats of biases are affecting the results increases.

Moving the P-value threshold from .05 to .005 will shift about one-third of the statistically significant results of past biomedical literature to the category of just “suggestive.”⁸ This shift is essential for those to believe (perhaps crudely) in black and white, significant or nonsignificant categorizations. For the vast majority of past observational research, the categorizations would be welcome. For example, medication reauthorization studies show that only few past claims from observational studies with $P < .05$ represent causal relationships.⁹ Thus, the proposed reduction in the level for declaring statistical significance may decrease mostly new noise with relatively little loss of valuable information. The reproduced text

REPRODUCIBILITY PROJECT
Cancer Biology

The Reproducibility Project: Cancer Biology is a collaboration between the Center for Open Science and the National Cancer Institute to independently replicate selected results from a substantial number of published cancer biology research articles.

RESEARCH

RESEARCH ARTICLE

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration[†]

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. **Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results. 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of**

nature

1,500 scientists lift the lid on reproducibility

Science sheds light on the 'crisis' rocking research.

Henry Jones

22 July 2016 | Corrected: 28 July 2016

Open Science Collaboration

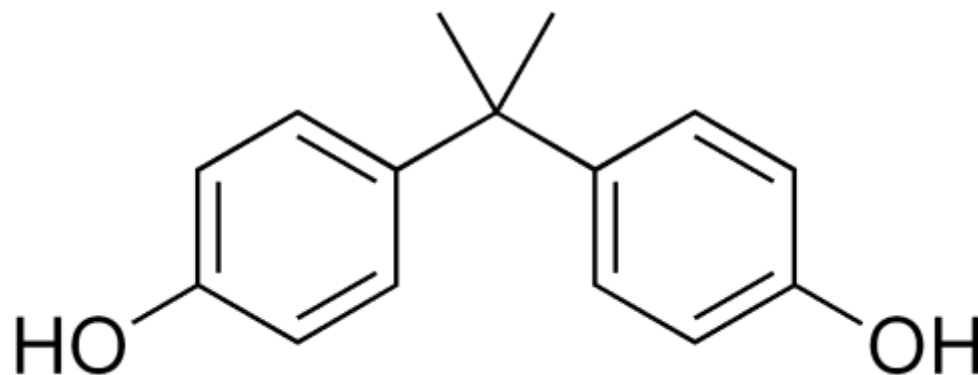
Science sheds light on the 'crisis' rocking research

Chemical risk assessment

- Making sense of complex and contradictory evidence about health risks posed by exposure to chemical substances

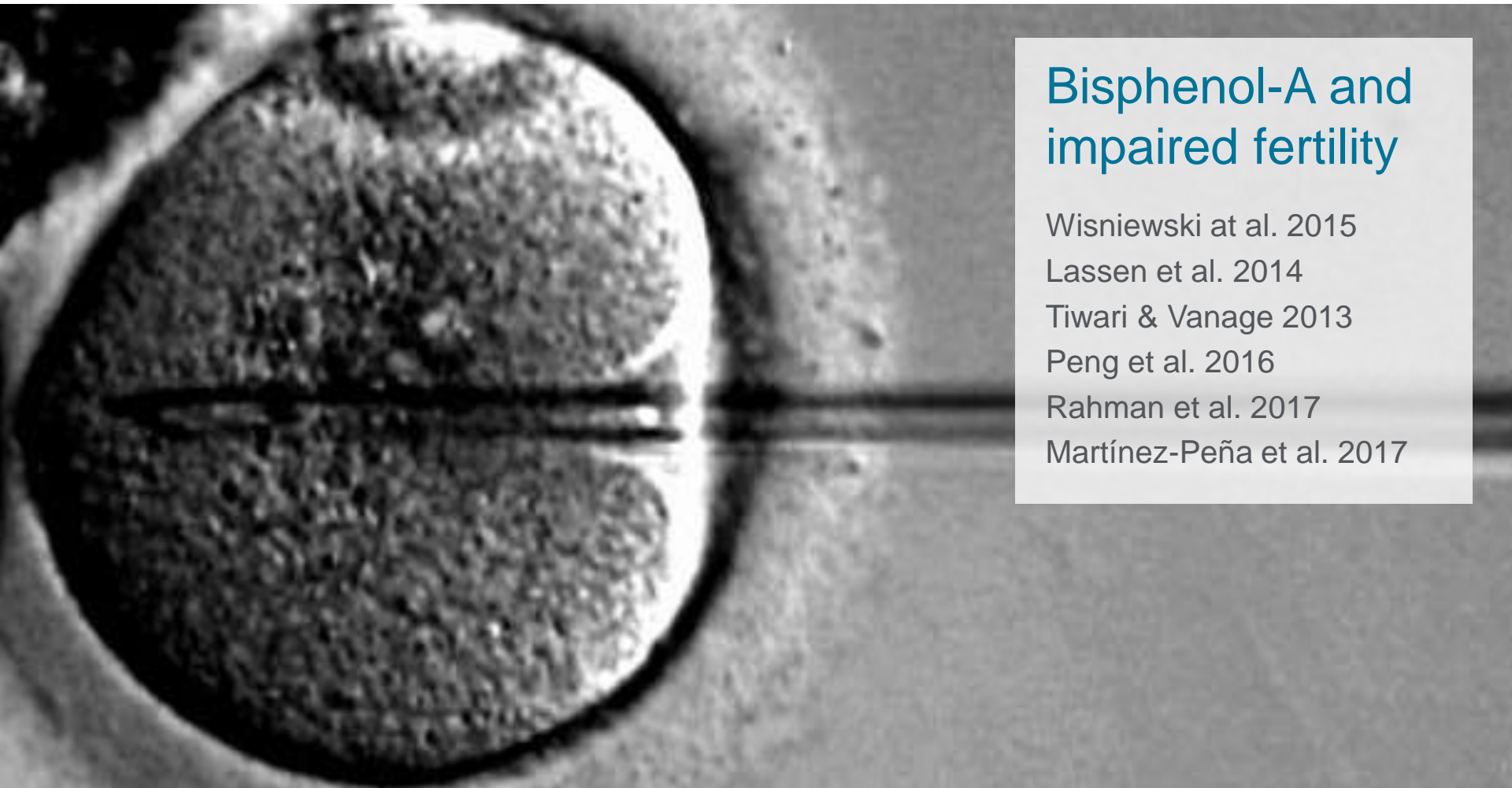


Reproducibility crisis in chemical risk assessment



Bisphenol-A





Bisphenol-A and impaired fertility

Wisniewski et al. 2015

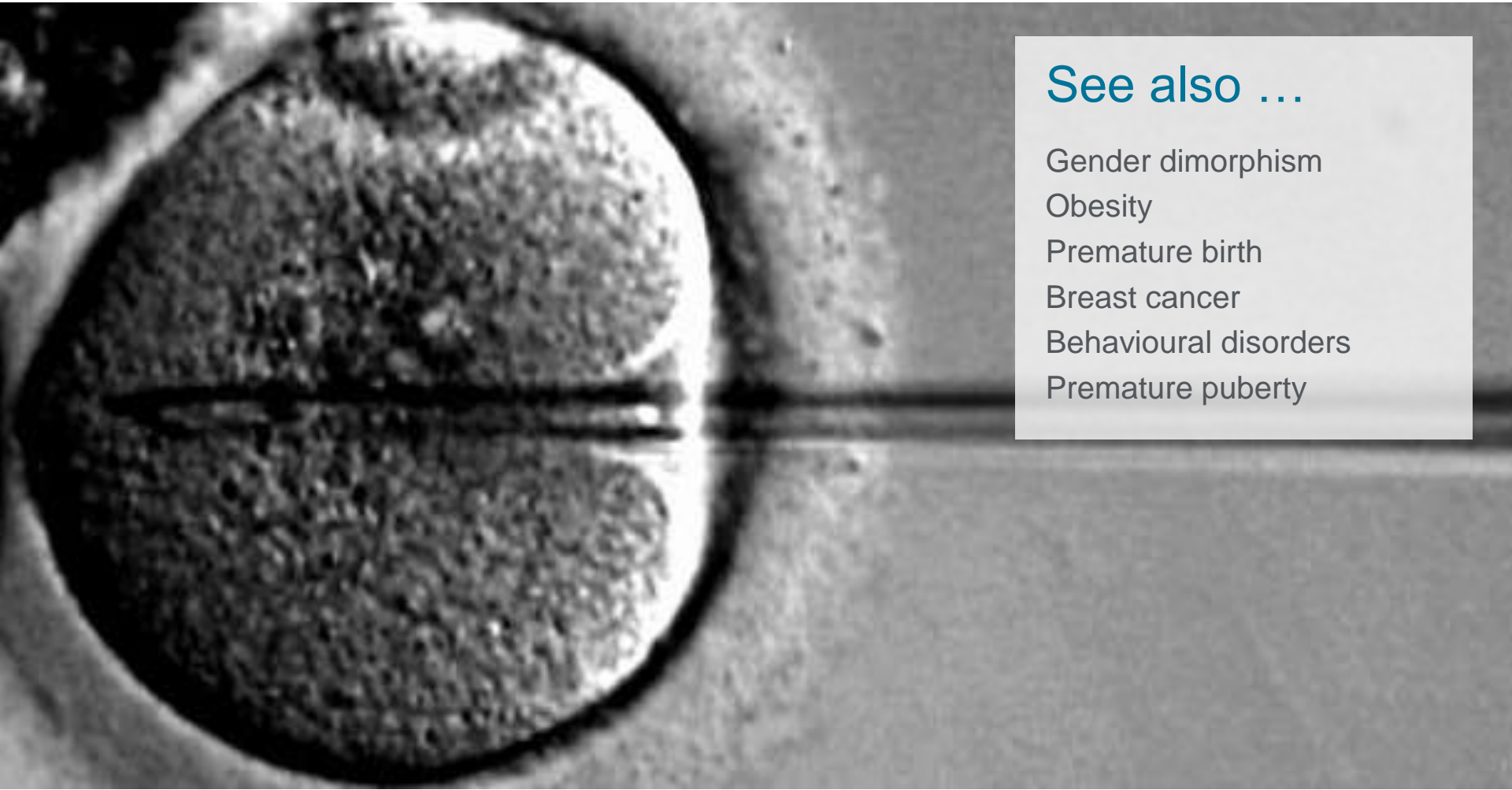
Lassen et al. 2014

Tiwari & Vanage 2013

Peng et al. 2016

Rahman et al. 2017

Martínez-Peña et al. 2017



See also ...

Gender dimorphism

Obesity

Premature birth

Breast cancer

Behavioural disorders

Premature puberty



Public Health
England



**Karolinska
Institutet**

**International Agency
Research on Cancer**



**World Health
Organization**

...effects have been demonstrated for BPA [at] levels **10–10,000x lower** than the current LOAEL of 50 mg/kg/day
[Vandenberg et al. 2014](#)

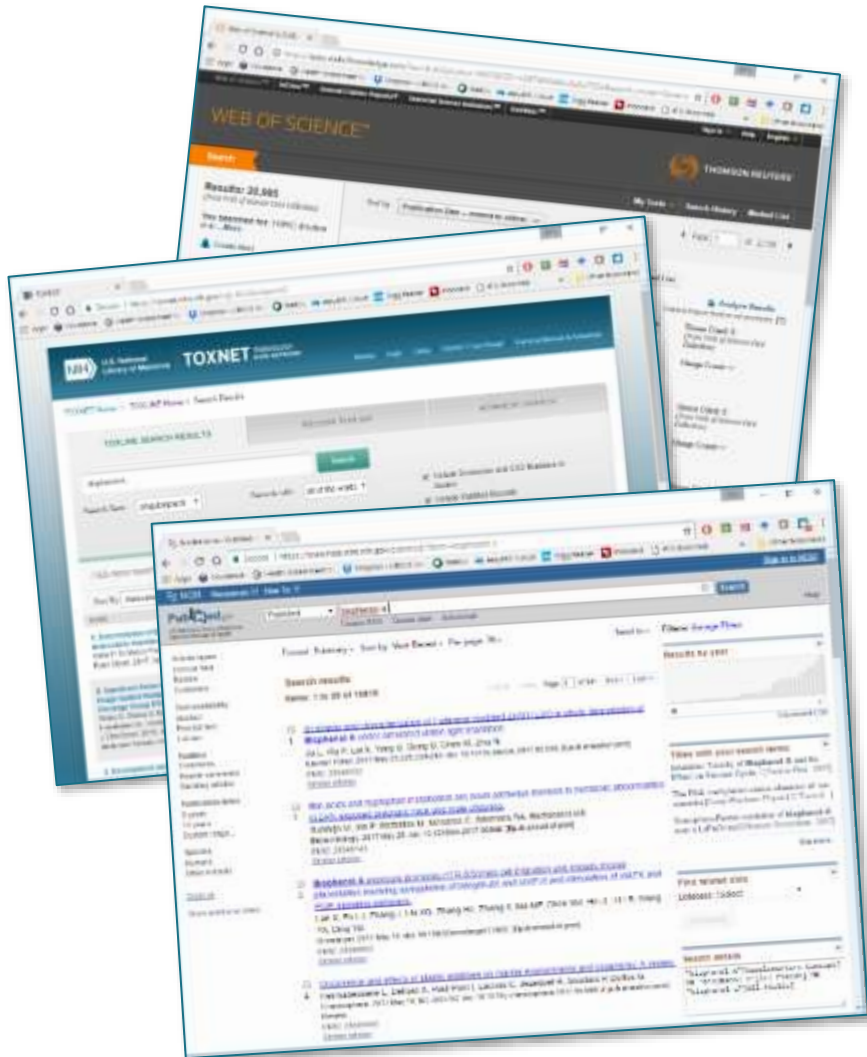


...**no health concern** for any age group from dietary exposure
[EFSA 2015](#)

...a TDI for BPA has to be **0.7 µg/kg bw/day** or lower to be sufficiently protective
[National Food Institute, Denmark 2015](#)

...a **potential risk to the unborn children** of exposed pregnant women [relating to] a change in the structure of the mammary gland
[ANSES 2013](#)

Same evidence, different conclusions



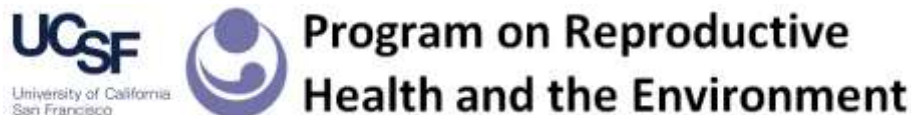
...no health concern for any age group from dietary exposure
EFSA

...a TDI for BPA has to be **0.7 µg/kg bw/day** or lower to be sufficiently protective
National Food Institute

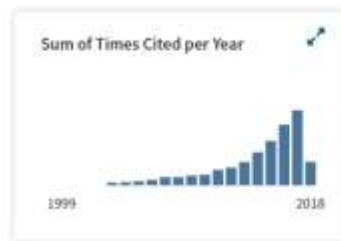
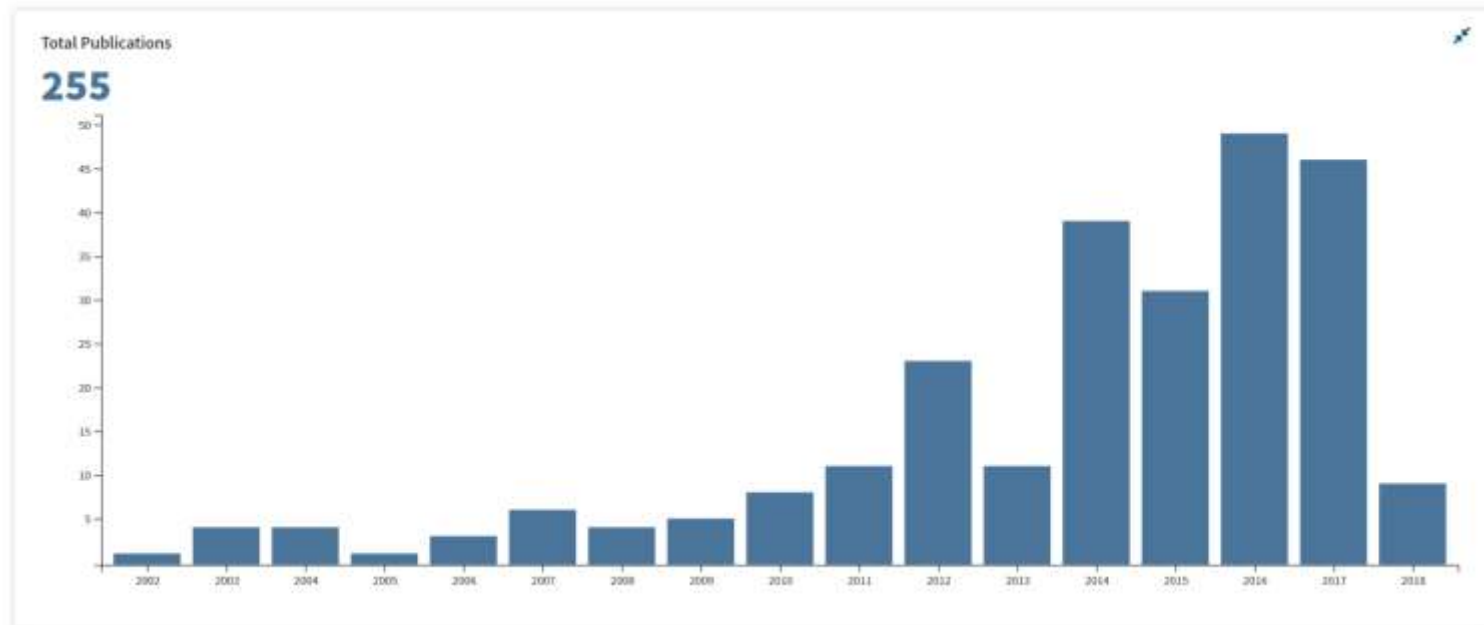
...effects have been demonstrated for BPA [at] levels **10–10,000x lower** than the current LOAEL of 50 mg/kg/day
Vandenberg et al. 2014

Solving the problem with systematic review methods

- Accelerating uptake since I started working on this in 2010



Rapid growth in publication of SRs



h-index 46

Average citations per item 22.6

Sum of Times Cited 5,763

Without self citations 5,683

Citing articles 5,306

Without self citations 5,260

TITLE: ("systematic review"); Refined by: WEB OF SCIENCE CATEGORIES: (TOXICOLOGY) AND [excluding] WEB OF SCIENCE CATEGORIES: (PHARMACOLOGY PHARMACY); Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, IC.

But we have a problem with quality

- 8989 PubMed records tagged by 2004 as “systematic review” yet actual number of stringently-defined SRs was ~2500 (Moher et al. 2007)
- Most published SRs have major flaws in conduct and reporting (Page et al. 2016)
- ~3% of manuscripts are “decent and clinically useful” (Ioannidis 2016)
- Our own pilot data shows serious omissions in reporting of 19 of 25 SRs published in the top environmental health journals through 2014-2015, before we even look at the validity of the actual methods used
- Fundamental errors mean a lot of effort is being put into projects which are not fit for purpose

My job as an editor

- What can I do at our journal to ensure each SR we publish is fit for purpose?
 - Asks an important question
 - Is truthful
 - Includes all information about methods and results, such that a reader can appraise the validity of the SR's findings and assess its relevance to their decision-making context
- Gatekeeper and midwife strategies for ensuring we publish high-quality research
- Implications for you as researchers

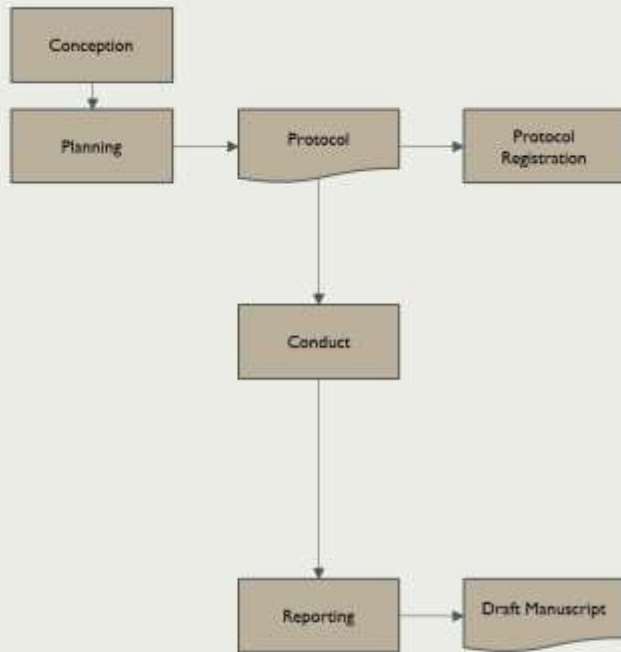
EDITOR AS GATEKEEPER

Enforcement of reporting standards

Editorial triage

Making best use of peer-review

Journal Side



Researcher Side

Enforcement of reporting standards

- Option of PRISMA (Moher et al. 2009) or ROSES (Haddaway et al. 2018)
- Submission of PRISMA or ROSES report as supplemental information is compulsory
- Useful quick check on basic standards

PRISMA Report (modified) for Systematic Maps Submitted to Environment International
Version 1.0, 29 Feb 2017. This form is to be completed on supplemental information alongside any systematic map submitted to Environment International. Authors are asked to provide reference numbers to address to page numbers.

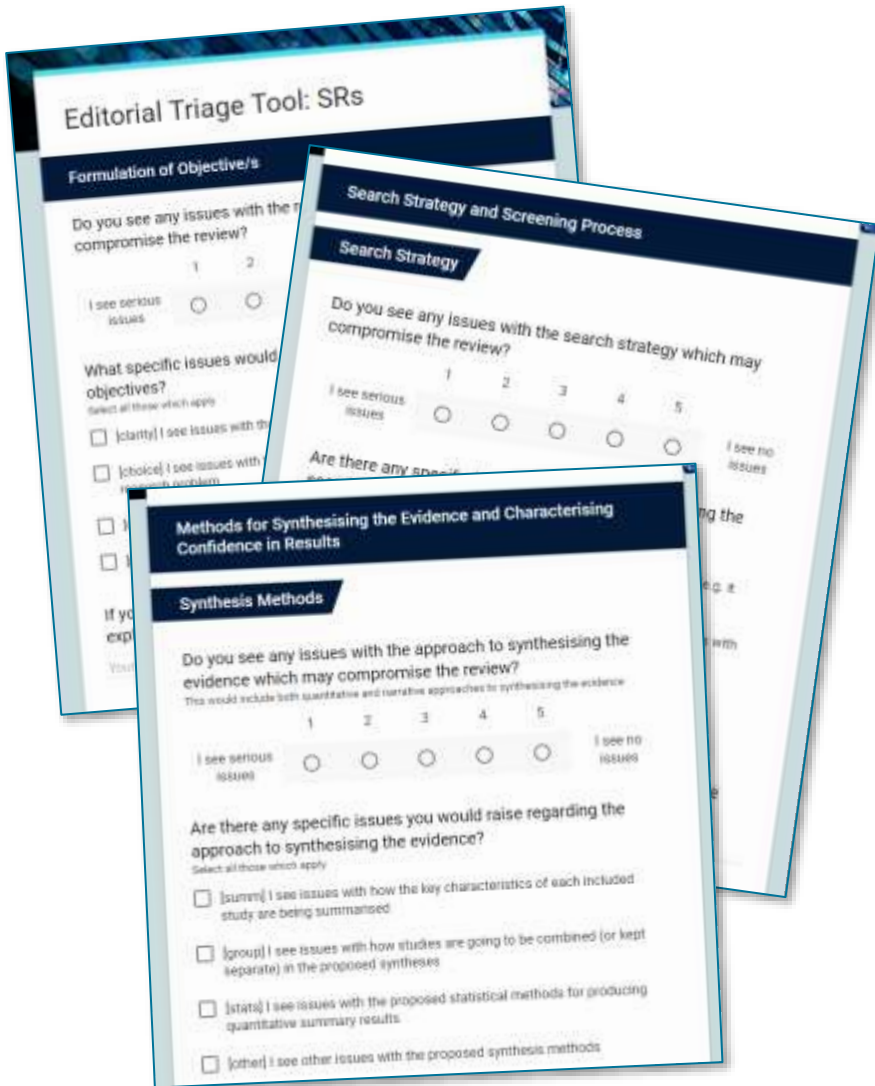
Title of submitted paper and corresponding author: (insert here)

#	Item	Guidance	On page #	Manuscript Quote / Author Comments
Title				
1	Title	Identify the report as a systematic map.		
Abstract				
2	Structured summary	Provide a structured summary including, as applicable: <ul style="list-style-type: none"> • Background; • Objectives; • Data sources; • Study eligibility criteria; • Study appraisal methods, if conducted; • Results; • Limitations; conclusions and implications of key findings; • Systematic map registration number. 		
Introduction				
3	Rationale	Describe the rationale for the map.		
4	Objectives	Define primary and secondary questions for the systematic map.		
Methods				
5	Protocol and registration	Indicate if a map protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.		
6	Eligibility criteria	Specify characteristics of study reports used as criteria for eligibility, giving rationale.		
7	Information sources	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.		
8	Search	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.		

The image shows the ROSES website interface. At the top, there is a navigation bar with the ROSES logo and the text "ROSES Reporting standards for Systematic Evidence Syntheses". Below the navigation bar, there are four main content areas, each with a "To the form" button:

- ROSES for Systematic review protocols
- ROSES for Systematic review reports
- ROSES for Systematic map protocols
- ROSES for Systematic map reports

Editorial triage reports



Environment International
Systematic Review Editorial Triage Report

Title of systematic review
[redacted] systematic review and meta-analysis

Name of lead author
[redacted]

Name of handling editor:
Paul Whaley

05/10/2018

1. Formulation of objectives

Reviewer satisfaction score (1 = serious concerns; 5 = no concerns)
2

Specific issues raised regarding the research objectives:
[clarity] I see issues with the clarity of the research objectives

Comments:
The objectives are not completely clear. While there is an intent to compare incidence of microbial contamination between bottled vs. mineral water, the importance of this particular comparison is unclear (why not just study prevalence of contamination, period, and see which subgroups of bottled water are at highest risk of contamination), and the significance of the connection to health effects which the authors emphasise is not apparent (is there a threshold level which contaminated bottled water crosses? If so, where? etc.). What counts as "contamination" is also not defined - is this a threshold level of microbiota, or mere presence?

2. Search strategy

Reviewer satisfaction score (1 = serious concerns; 5 = no concerns)
2

Specific issues raised regarding the search strategy:
[rep] There are issues with the reporting of the search strategy (e.g. it might not be reproducible), [miss] The search strategy will miss relevant evidence (e.g. issues with search strings, number of databases, etc.)

Comments:
The search strategy could be more clearly reported (e.g. in tables in supplemental information) than it is, as a narrative sequence in a paragraph in the main text. There is no obvious use of exploded search terms, while some seem either restrictive or redundant (e.g. searching "water" AND "bottled water"), which are a bit strange in terms of Boolean operator (why AND? and redundant; "Cluster" should contain

Improved peer-review

- Target of 4 reviewers per submission
 - 2 topic experts
 - 2 methods experts
- Peer-review facilitation tool
 - Testing a Google Forms tool similar to Triage tool
 - Building CREST-SR for full-blooded implementation

Whaley et al. "A Tool for Critical Appraisal of Evidence Syntheses in Toxicology: Systematic Reviews (CREST-SR)" Under development

1. Specifying review objectives

1.1 Rationale for the review
Appraisal target: evaluating whether the issue being addressed by the researchers is of sufficient importance to justify the conduct of a systematic review.

1.1.1 Rationale. Has the decision to conduct and publish a review been adequately justified?

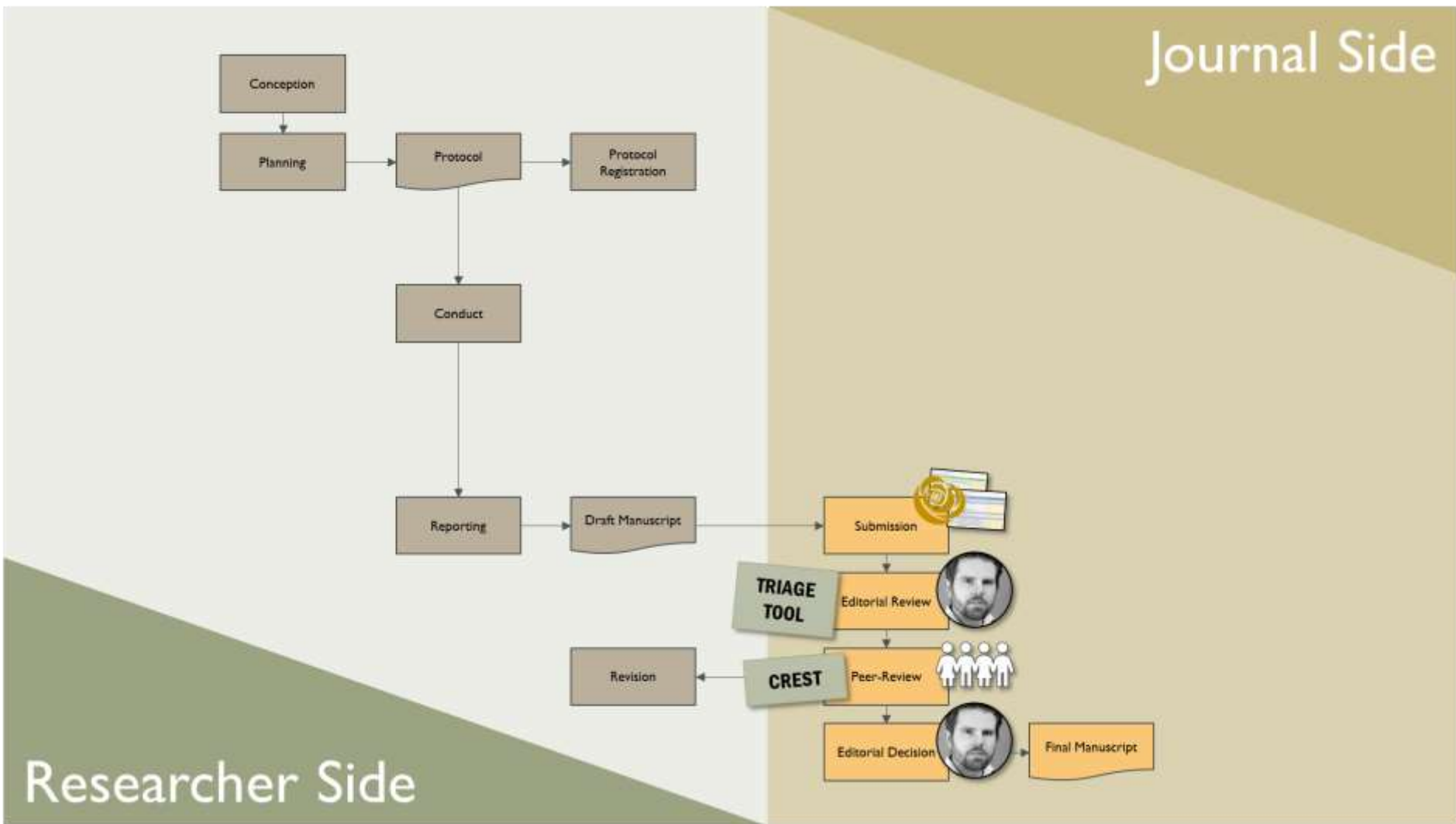
Level of concern:	<input type="checkbox"/> None	<input type="checkbox"/> None-Minor	<input type="checkbox"/> Minor	<input type="checkbox"/> Minor-Mod	<input type="checkbox"/> Moderate	<input type="checkbox"/> Mod-Major	<input type="checkbox"/> Major
--------------------------	-------------------------------	-------------------------------------	--------------------------------	------------------------------------	-----------------------------------	------------------------------------	--------------------------------

<p>Explanation:</p>	<p>Guidance points:</p> <ul style="list-style-type: none"> Resolves scientific uncertainty? Important to policy decisions? Important to stakeholders?
----------------------------	---

Recommendations for manuscript in relation to justification of conduct of the review

Can the concerns with the review as identified above be addressed by revising the manuscript?	<input type="checkbox"/> No concerns	<input type="checkbox"/> Yes	<input type="checkbox"/> No
If the concerns cannot be addressed via revisions, would the manuscript still be publishable if the shortcomings in the review were made clear to the reader?	<input type="checkbox"/> No concerns	<input type="checkbox"/> Yes	<input type="checkbox"/> No

Describe in appropriate detail any specific revisions and clarifications which need to be made to the manuscript:



Progress so far?

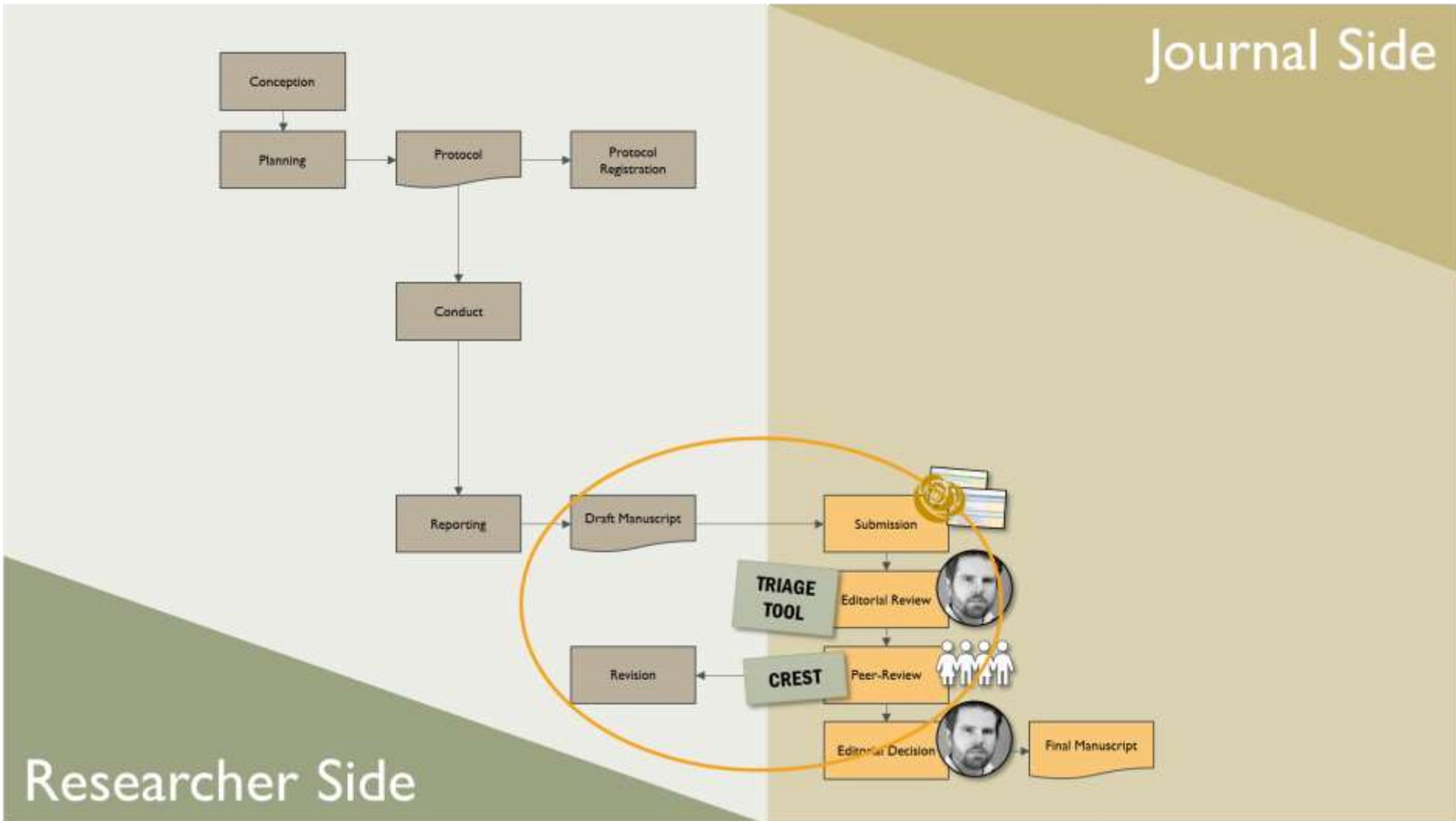
- 46 of 67 submissions rejected since using EVISE (~18 months)
 - 10 in process, 10 sent to production, one declined resubmission
 - 6 SRs, one SM, 2 commentaries, one correspondence
 - Only 3 SRs rejected post peer-review, 43 pre peer-review
- Hopefully that means we are at least filtering out the SRs which are not fit for purpose

Is it really progress?

- We are mainly getting low-quality systematic reviews long after it's too late for the authors to address major issues (43 of 46 rejections are at desk; 2 years of work rejected in 2 minutes)
 - Objectives lacking research value and/or focus
 - Insensitive search strategies
 - Inappropriate inclusion criteria
 - Inadequate or non-existent risk of bias assessment methods
 - Unstructured, unsystematic interpretation of strength of evidence
- We are making sure readers aren't receiving misleading research (at least through our own journal) but could do much more to help submitting authors develop high-quality manuscripts

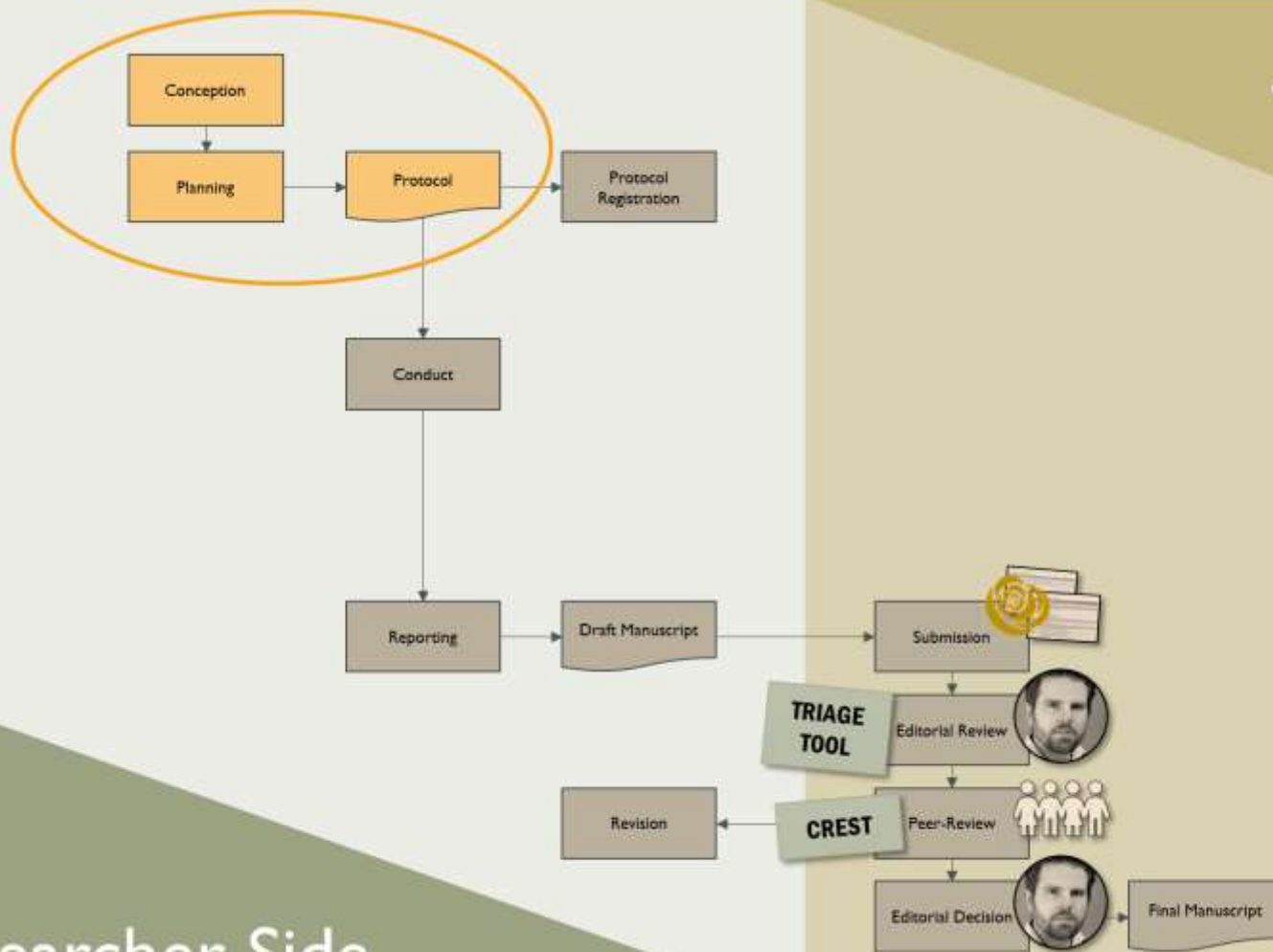
EDITOR AS MIDWIFE

Rethinking the SR workflow and submission process



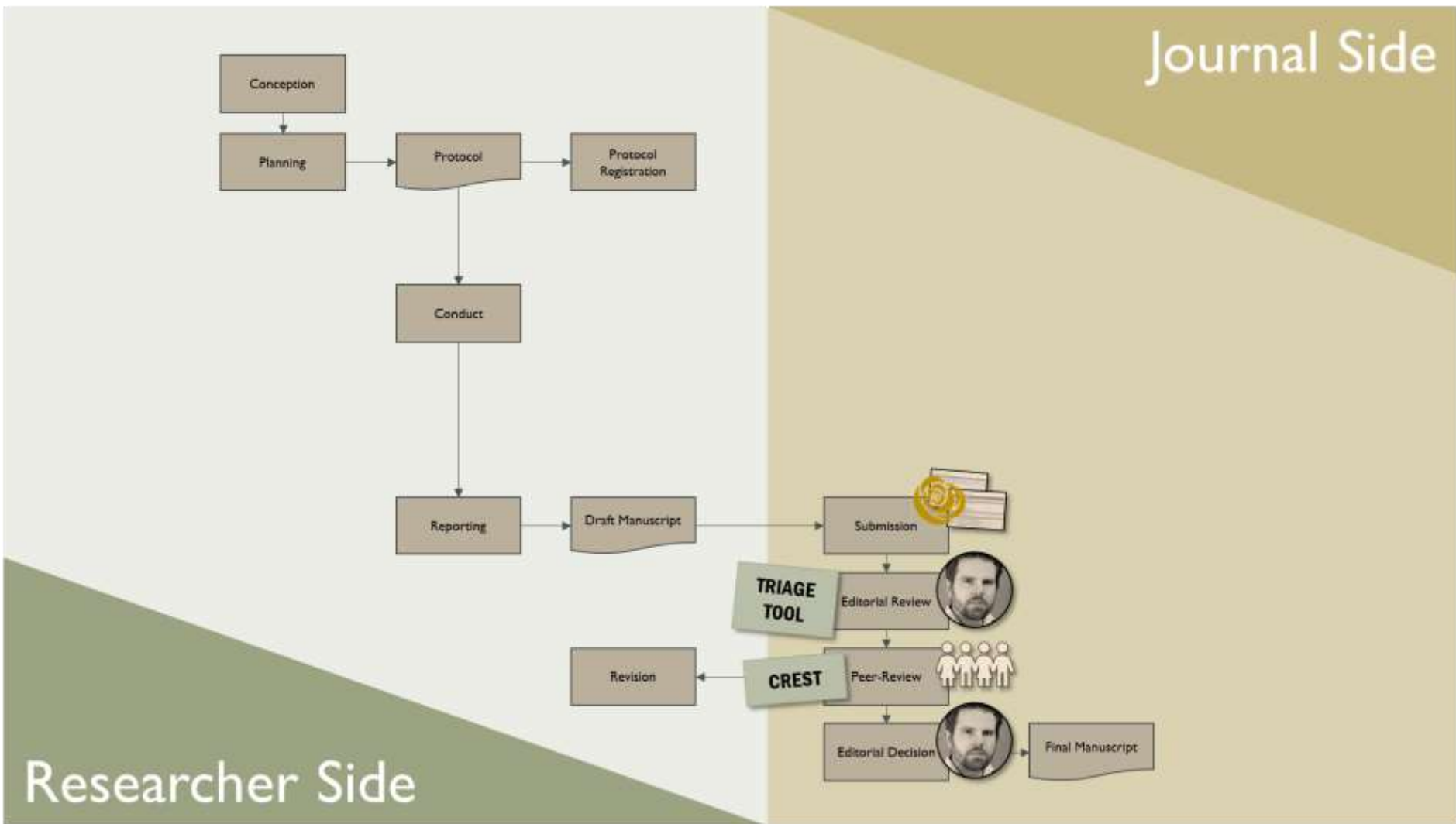
Journal Side

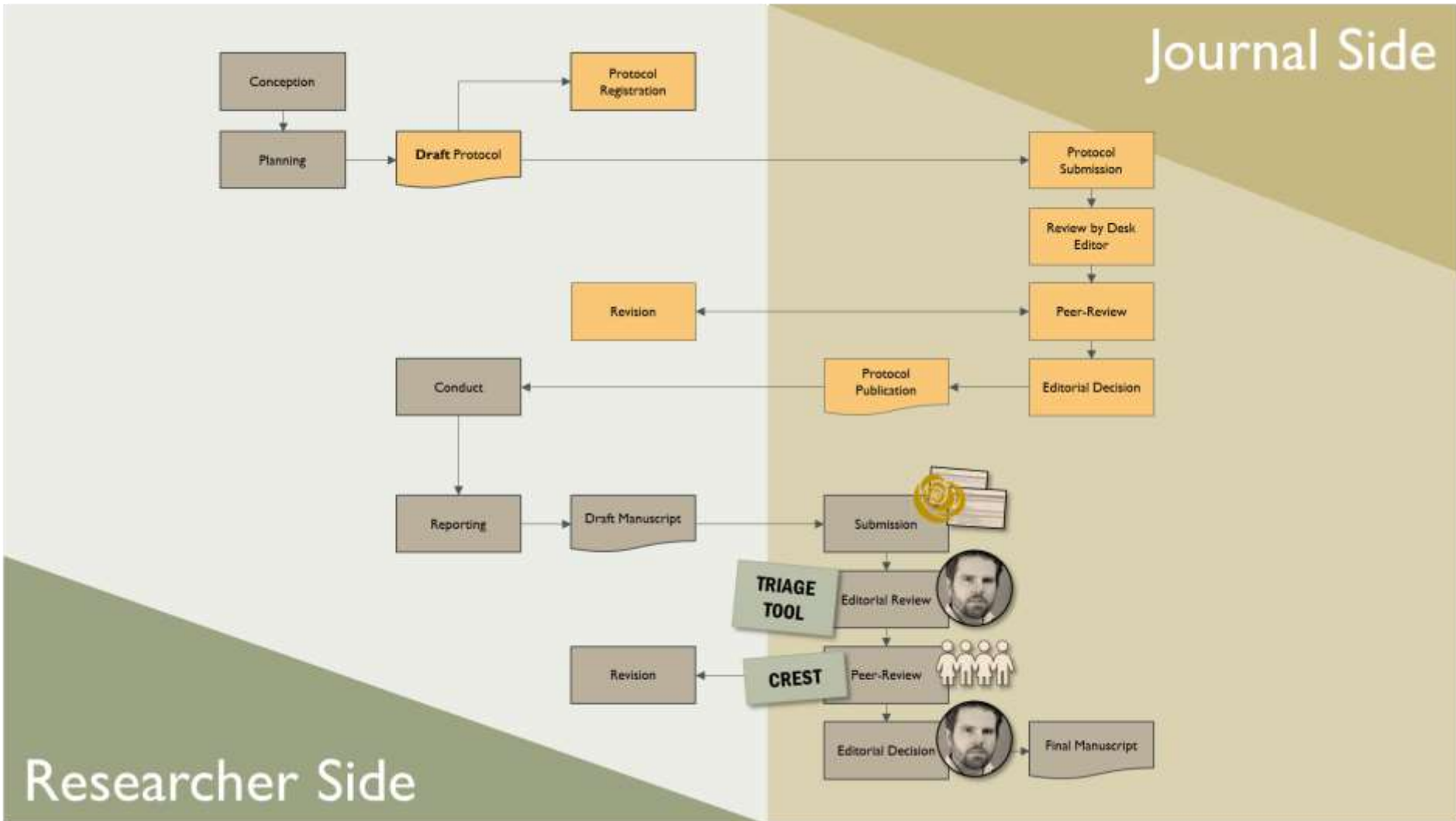
Researcher Side

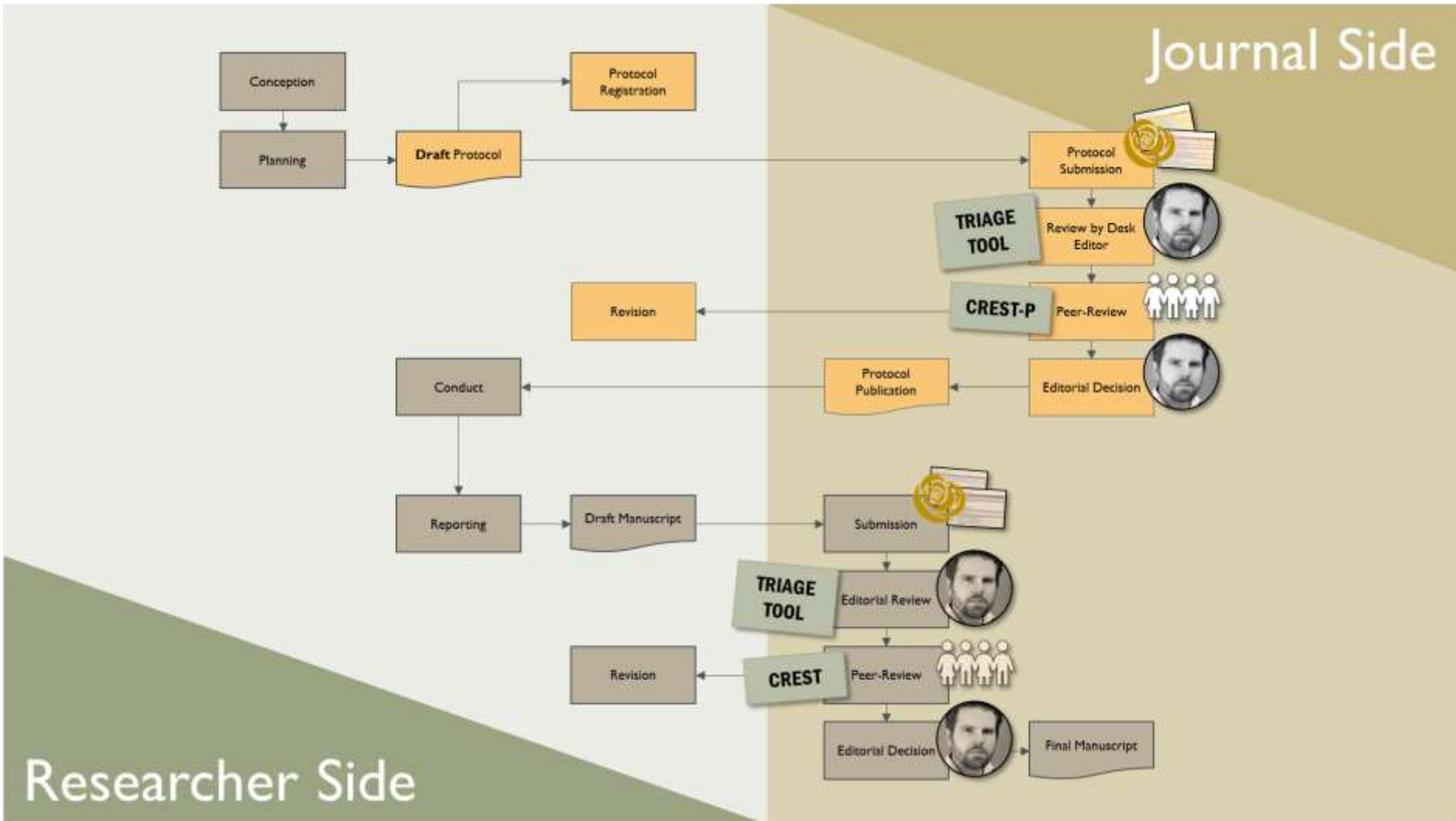


The solution: accept protocol submissions

- *Environment International* counts protocols as full publications
- First environmental health journal to do this
- Opens up multiple opportunities for editorial interventions





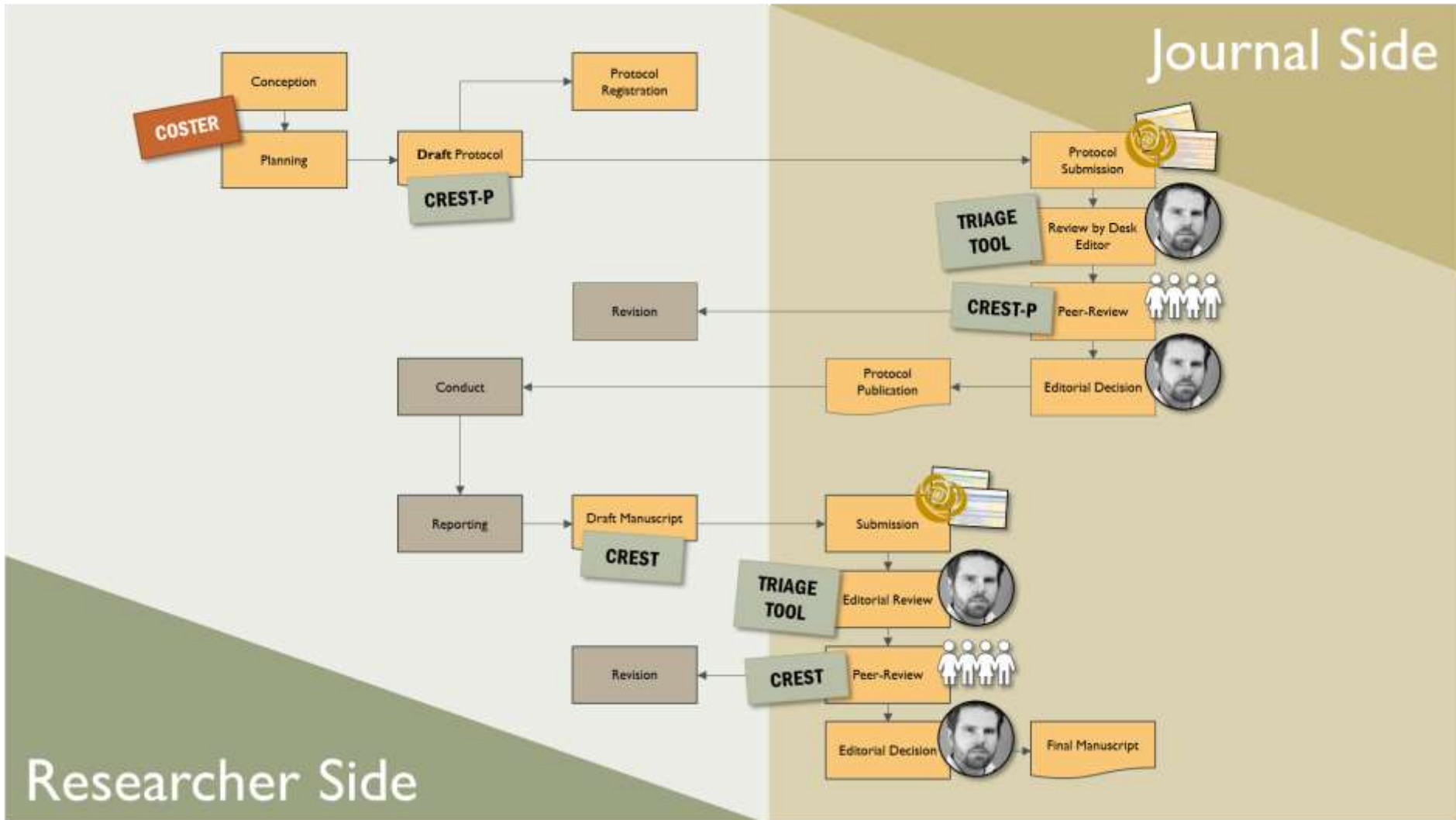


Final piece of the puzzle

- “Recipe-book” for what researchers ought to do, to maximise chance of producing a fit-for-purpose systematic review
- Developing a tool called COSTER – 70 provisions across 8 stages of conducting a systematic review
- Makes explicit the required processes for fulfilling the criteria of e.g. PRISMA or ROSES, and for critical appraisal tools such as CREST

Step 3: Screening Evidence for Inclusion

Proposed Wording	Comments	Notes for explanation / elucidation document
3.1 Screening of each piece of evidence for inclusion to be conducted by at least two people working independently, with an appropriate process (e.g. third party arbitration) for identifying and settling disputes.		
3.2 Document decisions in enough detail to allow presentation of the results of the screening process in a PRISMA flow chart.		



Implications for submitting authors

- Take advantage of our offer to review and publish protocols
- Follow best-practice standards for conduct of systematic reviews
- Think about the conduct implied by reporting standards
- For internal QC, use the same triage and peer-review tools we do
- Don't assume that any stage of a systematic review is optional
- It's good to be boring (results are irrelevant if methods are good)
- Find out more? **Subscribe to our newsletter:** <http://bit.ly/overcite>



overcite//
new developments in systematic review methods
for environmental health research

This month in **overcite// *** (scroll down)

New methodology publications: GRADE for assessing certainty in evidence from animal studies; guidance on gray literature searching; stakeholder engagement for controversial fields of regulatory science; exploring the concept of "WikiREACH"; evidence gap maps.

Issues in current SR practices: Pooled results of studies investigating adherence to the PRISMA Statement; prevalence of flawed statistical analyses in systematic reviews.

*Readers should note that all items are listed for interest only and not endorsed. Content subject to change.



new methodology publications//

GRADE // [Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies](#) The authors present how the GRADE approach could be used to rate certainty in the evidence from preclinical animal

Thank you.

Paul Whaley | p.whaley@lancaster.ac.uk



Image rights (in order of use)

Smokestacks / Guy Gorek / Flickr / CC BY-NC-ND

Plastic bottles / zhrefch / Flickr / public domain

Pesticide application / Oregon State University / Flickr / CC BY-SA

Food cans / King of Hearts / Wikimedia Commons

Till receipt / Till Dettmering / Wikimedia Commons

Drinks bottles / Amraepowell / Wikimedia Commons

Injected egg / Ekem / Wikimedia Commons